Gorter et al. BMC Medical Research Methodology (2015) 15:55
DOI 10.1186/s12874-015-0050-x

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Why item response theory should be used for longitudinal questionnaire data analysis in medical research

Rosalie Gorter[1,2*], Jean-Paul Fox[3] and Jos W. R. Twisk[1,2]

Abstract

Background: Multi-item questionnaires are important instruments for monitoring health in epidemiological longitudinal studies. Mostly sum-scores are used as a summary measure for these multi-item questionnaires. The objective of this study was to show the negative impact of using sum-score based longitudinal data analysis instead of Item Response Theory (IRT)-based plausible values.

Methods: In a simulation study (varying the number of items, sample size, and distribution of the outcomes) the parameter estimates resulting from both modeling techniques were compared to the true values. Next, the models were applied to an example dataset from the Amsterdam Growth and Health Longitudinal Study (AGHLS).

Results: The results show that using sum-scores leads to overestimation of the within person (repeated measurement) variance and underestimation of the between person variance.

Conclusions: We recommend using IRT-based plausible value techniques for analyzing repeatedly measured multi-item questionnaire data.

Keywords: Longitudinal data, Hierarchical model, Item response theory, Questionnaires, Measurement error, Structural model, Plausible values, Multilevel model

Background

In the field of medical epidemiological research, multi-item questionnaires are often used to measure the development of a subject's health status over time. The resulting item observations are used as measurements of a continuous latent variable (i.e. a variable that is not directly observable). Examples of latent variables are health related quality of life [1, 2], and depression [3]. A measurement model is required to describe the relation between the observed categorical item responses (for example, Likert items with four answering categories: agree/slightly agree/slightly disagree/disagree) and the continuous latent variable.

To make statistical inferences about longitudinal measurements of the latent variable a statistical model is required that describes the development of the latent variable over time, while addressing the typical correlations between measurements of one person. The central question is how to measure the latent variable given the response data, and how to perform the longitudinal data analysis given the measured variables. In longitudinal designs, the data has a nested structure; i.e. repeated measurements are nested within the subjects. Due to the nested structure, the common independence assumptions between measurements do not hold and neither linear/logistic regression nor analysis of variance can be used in a straightforward way [4–8]. A multilevel model can be used to model the dependencies when there are multiple measurements nested within participants [9]. This multilevel modelling approach will be referred to as structural modeling to explore differences in longitudinal analyses with sum-scores and IRT-based scores as estimates for the latent variable.

Two fundamental theoretical frameworks can be used to measure latent variables given the response data. Historically, there is classical test theory (CTT), where sum-scores are the estimates of the latent variable. The other, theoretically more advanced framework is item response

* Correspondence: r.gorter@vumc.nl
[1]Department of Epidemiology & Biostatistics, VU university medical center, Amsterdam, Netherlands
[2]EMGO+ institute for health and care research, Amsterdam, Netherlands
Full list of author information is available at the end of the article

theory (IRT) where item response patterns are used to construct scores for the latent variable. Under CTT, item differences are ignored and sum-scores have a common measurement error variance across subjects. Under IRT, different scores are assigned to the different response patterns leading to the same sum-score, making it possible to distinguish between the latent variable scores of subjects with similar sum-scores. Item response patterns are lower-level observations and more informative about the latent variable than higher-level aggregated sum-scores, which ignore differences between response patterns leading to equal sum-scores. Another advantage of IRT is that the distribution of the latent variable can address skewness of the population distribution, where under CTT, the distribution of the latent variable is restricted to be symmetric. In most epidemiological studies however, a symmetric latent variable population distribution is not present [10, 11]. Despite the known benefits of IRT, epidemiological researchers are still using sum-scores [12–15] as estimators of the latent variable.

The measurement error associated with latent variables is usually ignored in the structural model when using sum-scores with equal amounts of measurement error for all scores on the latent variable. The parameter estimates of the structural model will be biased [16] consequently. To address the uncertainty associated with the measurements, the plausible value technology [17–21] can be used. In plausible value technology, several draws (mostly five [22]) from the posterior distributions of latent variable scores for each person are used as latent variables in the structural model. The results from the structural model are pooled for all draws to obtain parameter estimates. Plausible value technology can be used to address directly the negative implications of using sum-scores as measurements of latent variables, while making the comparison with IRT-based plausible values.

The objective of this paper is to stress the important differences between IRT and CTT for latent variable modeling in different situations and show why IRT measurement models should be used in longitudinal research. As a case study in epidemiological longitudinal data, the repeatedly measured trait anxiety questionnaire from the Amsterdam Growth and Health Longitudinal Study (AGHLS) [23] is used.

Methods

Structural model for longitudinal latent variables

A structural model (also known as latent regression model) describes the relationships between predictors and latent variables while addressing additional dependencies between the repeatedly measured latent variables. A well-known method to account for the nested structure of longitudinal data is multilevel modeling (or mixed modelling, random effects modelling, hierarchical linear modelling) as the structural model. Advantages of using multilevel modelling for longitudinal data analysis are that it is not necessary that subjects are measured on the same time points nor do follow up times need to be uniform. Furthermore, the model is capable of handling time-invariant and time-variant covariates. Also, it is possible to estimate subject-specific change across time. The following multilevel model will be considered,

$$\begin{aligned} \theta_{ij} &= \beta_j + e_{ij} \\ \beta_j &= \gamma + u_j \end{aligned} \tag{1}$$

where θ_{ij} is the latent variable location of person j for measurement occasion i, β_j is the random intercept representing the average latent variable location of person j over measurement occasions; both error terms are normally distributed with $e_{ij} \sim N(0, \sigma^2)$, and $u_j \sim N(0, \tau^2)$. The variance parameter τ^2 of the multilevel model is the variance between persons (i.e. level-2 variance) whereas σ^2 is the repeated measurement variance (i.e. the variance of the measurements within person; level-1 variance).

Measurement part of the model

To estimate the latent variable θ_{ij} that is used in the structural model as described in equation Eq. 1, CTT or IRT-based methods can be used. Lord and Novick [24] pp. 44 describe the basic equation for the composition of the observed score, X_{gij}, for the latent variable, T_{gij}, for person j on measurement occasion i on questionnaire g as

$$X_{gij} = T_{gij} + E_{gij}, \tag{2}$$

where T_{gij} is the true score, and E_{gij} the error of measurement. The observed score consists of the true score and the error of measurement, which is assumed to be unbiased. When making this assumption about the measurement error, numbers can be attached to the answering categories of the items and summed over all items of the questionnaire. Then, a test score (i.e., sum-score) can be defined as

$$\theta_{ij_cTT} = \sum_{k=1}^{K} X_{kij}, \tag{3}$$

where the response pattern for person j on measurement occasion i is given by (X_{1ij}, …, X_{Kij}), and where K represents the number of items in the questionnaire. These sum-scores are the CTT estimates for the latent variable and used as outcome variable in the longitudinal analysis (i.e. the structural model). There are two main problems with this way of quantifying latent variables.

The first issue is that the characteristics of the test and the subject are inseparable, i.e. they cannot be interpreted without the other, which makes sum-scores population dependent. The second problem is that the standard error of measurement is assumed to be the same for all subjects, although some sum-scores are more informative about the latent variable than others. That is, it is much more likely that different subjects are measured with different precision. For example, extreme high or low sum-scores are more unreliable compared to average sums scores, meaning that the extreme sum-scores are less likely to distinguish between the subjects than the sum-scores in the middle of the scale.

The item response patterns are more informative about the latent variable than the aggregated sum-scores, which ignore differences between response patterns leading to the same sum-score. For example, when answering 'yes' to 10 out of 20 dichotomous items (1 = '*yes*', 0 = '*no*'), the sum-score of 10 can be obtained in 20!/10! ways. Under IRT different scores are assigned to the different response patterns all leading to the same sum-score, making it possible to distinguish the scores of respondents with similar sum-scores.

Using IRT modeling is an accepted way to account for the differences in measurement precision between persons [25–29] and can be used to estimate scores for the latent variable. In IRT, the relation between the unobserved latent variable θ and the observed item scores are described by item characteristic curves that model the probability of observed item responses. As a result, the item and latent variable estimates in IRT modeling are not dependent upon the population [30]. Fig. 1 depicts an example of item characteristic curves for an item with four response categories where the probabilities of choosing a certain category are plotted against the latent variable. An IRT model describes the relationship between latent variables and the answers of the persons on the items of the questionnaire measuring the latent variable [31]. For ordered response data, the probability that an individual indexed *ij* with an underlying latent variable θ_{ij}, responds into category c ($c = 1, \ldots, C$) on item k is represented by

$$p\left(\boldsymbol{y}_{ijk} = c|\theta_{ij}, a_k, \tau_k\right) = \phi\left(a_k\theta_{ij} - \tau_{kc-1}\right) - \phi\left(a_k\theta_{ij} - \tau_{kc}\right), \tag{4}$$

where τ_{kc} are the $C_k - 1$ threshold parameters. The probability that the response $\boldsymbol{y}_{ijk}$ falls into category c is the difference of the probability densities (ϕ) of category $c - 1$ and category c. The response categories are ordered as $-\infty \leq \tau_{k1} \leq \tau_{k2} \leq \tau_{k3} \leq \infty$.

This item response model is called the graded response model [32] (or ordinal probit model [31, 33]). A Rasch [34] restriction was used fixing the discrimination parameters, a_k, to one.

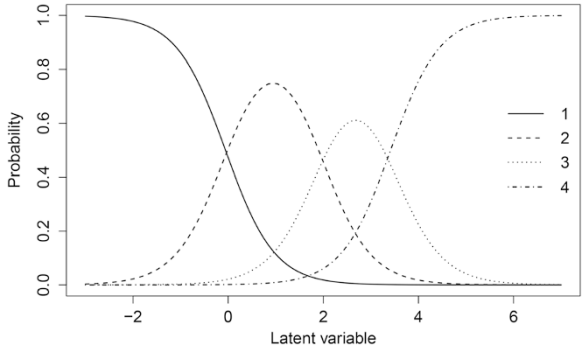

Fig. 1 Item response characteristic curves for item 2 from the STAI-DY with four answering categories. The item that was used for this example was 'I feel nervous and restless' with four answering categories '1. Almost never', '2. Sometimes', '3. Often', and '4. Almost always'. The crossing of two lines mark a threshold and can be interpreted as the location on the latent parameter where the probability of choosing the corresponding category or higher is 0.5

Computing and generating IRT- and CTT-based scores

Different methods exist for generating values for the latent variable in the IRT framework. The first way is to generate point estimates of the latent variable modeled by the IRT model by constructing the posterior distribution of the latent variable given the data. An important assumption of IRT modeling is conditional independence, which entails that response probabilities for items rely solely on the latent variable, θ_{ij}, and the item parameters. As a result, the joint probability of a response pattern $\boldsymbol{y}_{ij}$, given the latent variable θ_{ij}, of a person j on measurement occasion i, and given the item parameters, over the K items of the questionnaire is the product of the probabilities of the individual answers of a person on all items of the questionnaire given this person's position on the latent variable θ_{ij} (equation Eq. 5).

$$\begin{aligned} p\left(\boldsymbol{y}_{ij}|\theta_{ij}\right) &= p\left(\boldsymbol{y}_{ij1}|\theta_{ij}\right)p\left(\boldsymbol{y}_{ij2}|\theta_{ij}\right)\ldots p\left(\boldsymbol{y}_{ijK}|\theta_{ij}\right) \\ &= \prod_{k=1}^{K} p\left(\boldsymbol{y}_{ijk}|\theta_{ij}\right), \end{aligned} \tag{5}$$

When assuming a prior distribution for the latent variable distribution, $g(\theta_{ij})$, a posterior mean can be derived from the posterior $p(\theta_{ij}|\boldsymbol{y}_{ij}) = p(\boldsymbol{y}_{ij}|\theta_{ij})g(\theta_{ij})/p(\boldsymbol{y}_{ij})$, where the posterior is derived according to Bayes' rule [35]. The posterior mean can be used as an estimate of the latent score.

When using the posterior mean as an estimate of the latent score and thus as an outcome variable in the structural model, the uncertainty associated with the score is ignored. To account for this uncertainty, plausible value technology is used [18, 19, 36, 37] where the

latent outcome variable is treated as missing data. Plausible values are generated from the posterior distribution of the latent variable to obtain a complete data set. This data set can be analyzed in the secondary data analysis. When constructing the posterior of the latent variable, all available information is used. The posterior is proportional to the likelihood times the prior, which can be represented by

$$p\left(\theta_{ij}|\mathbf{y}_{ij},\sigma^2,\tau^2,\boldsymbol{\gamma}\right)\propto p\left(\mathbf{y}_{ij}|\theta_{ij}\right)p\left(\theta_{ij}|\sigma^2,\tau^2,\boldsymbol{\gamma}\right). \quad (6)$$

The structural model parameters are integrated out such that the (marginal) posterior of the latent variable only depends on the response pattern. This marginal posterior is only dependent on the data, in this case upon the data of subject j on occasion i. We sample from the marginal distribution in order to obtain plausible scores for subjects with similar response patterns and background characteristics as in the sample of subjects [20, 21].

For the CTT model, the sum score defined in equation Eq. 3 is considered to be an unbiased estimate of the true score. This true score is considered to be an outcome of the multilevel model in equation Eq. 1. When assuming the CTT model for the measurement of the construct score, according to equation Eq. 2, the distribution of the observed scores given the true score is given by $p(\mathbf{y}_{ij}|\theta_{ij\_CTT})$. Subsequently, the posterior distribution of the true score is given by

$$p\left(\theta_{ij_CTT}\middle|\mathbf{y}_{ij},\sigma^2,\tau^2,\gamma\right)\propto p\left(\mathbf{y}_{ij}|\theta_{ij_CTT}\right)p\left(\theta_{ij_CTT}|\sigma^2,,\tau^2,\gamma\right). \quad (7)$$

Parallel measurements are needed to estimate the true score (error) variance, but they are usually not available. When the measurement error variance cannot be estimated under the CTT model, the first term on the right-hand side is not included in defining the posterior distribution, and an unbiased estimate of the true score is assumed. However, the measurement error can still be assumed to be included in the multilevel model specification (i.e., the second term on the right-hand side). In that case, the population variance is used as an approximation of the subject-specific measurement error variance [24] pp. 155. This approach was also used in the present study.

Analogue to generating plausible values under the IRT model (equation Eq. 6), the marginal distribution of the (true) scores is used to generate plausible values under the CTT model. Note that the drawn plausible values are realizations of the true score under the structural multilevel model given the sum score as an unbiased estimate of the true score.

In the literature, it is recommended to draw five sets of plausible values to address the uncertainty associated with the plausible values for the missing data [37, 38]. Various results of data analysis are obtained for the five different complete data sets, which are constructed from multiple sets of plausible values. The final results are constructed by averaging the analysis results, in this case, the structural model from equation Eq. 1.

Comparing CTT and IRT-based estimates

When comparing the IRT and CTT-based structural model estimates, it is required to take scale differences into account. For the comparison, the CTT scores were rescaled to the IRT-based plausible values scale, using a linear transformation as proposed by Kolen and Brennan [39] pp. 337,

$$sc(y)=\frac{\sigma(pv)}{\sigma(Y)}y+\left[\mu(pv)-\frac{\sigma(pv)}{\sigma(Y)}\mu(Y)\right], \quad (8)$$

where $\mu(pv)$ and $\sigma(pv)$ are the mean and the standard deviation of the IRT-based plausible values, and where $\mu(Y)$, and $\sigma(Y)$ are the mean and standard deviation of the CTT-based plausible values.

Next, the structural model was fit to the plausible values for each of the five draws. Finally, the estimates resulting from the structural model were pooled to obtain the final parameter estimates.

Simulation study

A simulation study is presented for evaluating the use of IRT-based plausible values compared with CTT modeling for estimating latent variables used in longitudinal multilevel analysis. The aim of this study was to investigate how the true values of the population parameters are retrieved in different situations with varying sample sizes, number of items and skewness of the latent variable distribution.

Design

The full cross classified design resulting in 546 conditions is depicted in Table 1. Per condition, 10 datasets were simulated using R statistical software [40] and analyzed using an extended version of the R-Package mlirt [31] and WinBUGS [41, 42]. Data was generated following the model described in equation Eq. 1 with the variance between persons (i.e. level-2 variance) set to τ^2=.8, and repeated measurement variance to σ^2=.4 while using the IRT-model from equation Eq. 4 to generate values for the normally distributed underlying latent variable; $\theta_{ij} \sim N(0,1)$. An unidimentional latent variable was assumed to cause the responses, measured six times $J = 6$ using a varying amount of items with four answering categories per item $C = 4$. The number of items that

Table 1 Simulation conditions. The conditions of the full cross classified design for the simulation study. The numbers of items, number of participants, as well as the skewness from a normal distribution of the latent variable were varied

Model[a]	Items[b]	N[c]	Skewness
IRT	3	100	0
CTT	5	500	+/− 0.4
	7	1000	+/− 0.8
	10		+/− 1.3
	15		+/− 1.9
	20		+/− 2.6
	50		+/− 3.6

[a]Measurement model that was used for drawing the plausible values for the latent variable
[b]Number of items
[c]Number of participants

were used are listed in Table 1. In the simulated data, skewness of the latent variable was generated by changing the location of threshold parameters of the IRT model. For example, for a positive skewness of 2.6, $\tau_{k1} = 3$, $\tau_{k2} = 2$, and $\tau_{k3} = 1$ were used for all items k. This skewed to the right data could indicate that relatively healthy persons were asked to answer a questionnaire measuring clinical depression, leading to high sum-scores. The same can occur when subjects have recovered after treatment and the same questionnaire is used on the baseline and follow up measurement. Item scores were generated based on the latent variable and the parameters of the IRT model. The CTT based scores are calculated according to equation Eq. 3 using the simulated item scores. These sum-scores are the CTT estimates for the latent variable. In Fig. 2, the distributions of the CTT scores are visualized using density plots of different skewness conditions. In epidemiological data, skewness in the data is often found. The influence of various levels of skewness of the latent variable distribution on retrieving the multilevel regression parameters was investigated.

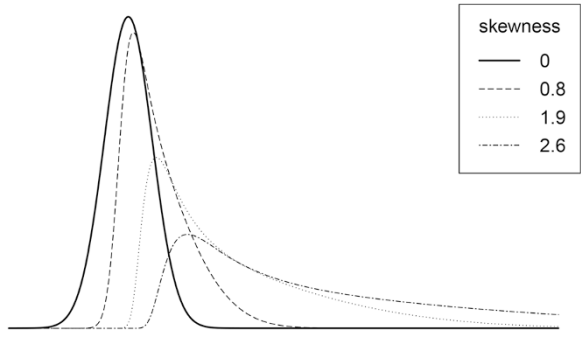

Fig. 2 The density of four different skewed normal distributions for the latent variable

Figure 3 shows a schematic display of the simulation procedure for one replication. IRT and CTT-based plausible values were generated using the datasets from one of the simulation conditions. The CTT-based Plausible values were rescaled according to equation Eq. 7 and the structural model described in equation Eq. 1 was fit to five draws of plausible values and the results were pooled by averaging the parameter estimates. Mean squared errors (MSE) given by

$$MSE\left(\hat{\theta}\right) = Var\left(\hat{\theta}\right) + \left(Bias\left(\hat{\theta}, \theta\right)\right)^2, \quad (9)$$

were calculated where $\hat{\theta}$ denote the parameter estimates resulting from the different replications and θ are the true values.

The MSE's were calculated for the level-1 and level-2 variance estimates for both the CTT and the IRT-based plausible value analysis.

Simulation Results

Figure 4 shows a selection of the variance estimates within persons (level-1) and between persons (level-2). The estimates for the IRT-based plausible value scores are closer to the true parameter value of 0.4 for the within person variance and 0.8 for the between person variance compared to the estimates from the CTT-based analysis. The difference between the methods is the smallest when the latent variable is perfectly normal distributed, and becomes gradually bigger with increasing skewness of the latent variable distribution. The estimates from the CTT model get closer to the true values when the number of items increase moving from the left to the right graphs. The estimates from the IRT model are close to the true values for number of item conditions with except for the $N = 100$ condition. The CTT repeated measurement variance estimates for the conditions with ten or less items are even higher compared to the between person variance estimates in case of more extreme skewness. This in contrary to the IRT-based estimates, where the variance estimates are very close to the true values regardless of the skewness. Overall, the IRT method gives more accurate estimates compared to the CTT model over all the simulation conditions. When comparing the plots from the top to bottom, the sample size increases from $N = 100$ to $N = 500$ to $N = 1,000$. Increase in sample size does not influence the magnitude of the difference between both models. The IRT model gives better estimates in all sample size conditions. With the increasing sample size, the lines between the estimates become more stable, indicating a more stable pattern of the differences between both methods. The MSE's for the variance estimates are presented in Fig. 5, where it can be seen that the CTT model

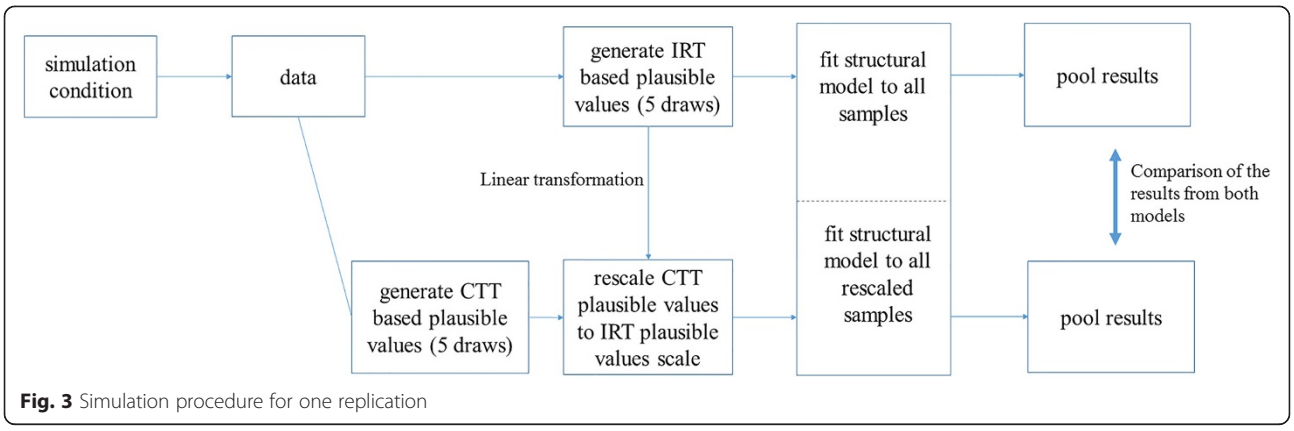

Fig. 3 Simulation procedure for one replication

systematically overestimates the repeated measurement variance and underestimated the between person variance. The differences between the true value and the estimated value by the CTT model increases when the latent variable distribution becomes more skewed. These differences become smaller for the CTT-based estimates when the number of items and the sample size increase. The observed difference between the IRT and CTT estimates seems to be dependent on the manipulated factors. The extremer data situations are causing larger differences between IRT and CTT-based estimates in a consistent way. A complete representation of the results can be found in Additional file 1 and Additional file 2 online.

Empirical dataset

An example dataset was analyzed to illustrate the application of IRT-based plausible values in epidemiological practice. Data were obtained from the Amsterdam Growth and Health Longitudinal Study (AGHLS), which is a multidisciplinary longitudinal cohort study that was originally set up to examine growth and health among teenagers [43]. Data from the AGHLS were used in previous research to answer various research questions dealing with the relationships between anthropometry [44], physical activity [45], cardiovascular disease risk [46, 47], lifestyle [48, 49], musculoskeletal health, psychological health [50] and wellbeing. The presented sample consists of 443 participants who were followed over the period 1993–2006 with maximal three data points over time nested within the individuals. A subscale of the State Trait Anxiety Index Dutch Y-version (STAI-DY) [51] questionnaire was used to measure the latent variable 'state anxiety' and consists of 20 items with four answering categories. The histograms in Fig. 6 depict the sum-score distributions on the three measurement occasions. The aim of the analysis was to estimate the intercept and the variance parameters (i.e. an intercept only model) in order to compare the CTT and IRT-based estimates. The measurement models as well as the structural model that were used are comparable to those in the simulation study above.

Results

The pooled parameter estimates resulting from both the IRT and CTT-based models are presented in Table 2. The parameter estimates are derived by pooling the averages from the posterior distributions of the five draws of plausible values that are visualized in Fig. 7. Looking at the estimate for the random intercept on the first row of Table 2, it can be seen that the IRT and CTT-based estimates are similar. Furthermore, it can be seen that the between person variance is lower for the CTT-based model compared to estimates from the IRT-based model, 0.685 and 0.733 respectively. The within person variance was 0.357 for the CTT-based model while it was 0.294 for the IRT-based model. As a result, the intra class correlation coefficient (ICC) was higher for the CTT-based model indicating that there was relatively more residual variance relative to the total variance compared to the IRT-based model. As a result, the CTT-based model overestimates the ICC substantially. Looking at the posterior distributions resulting from the first draw of plausible values on the top row of plots in Fig. 7, the two distributions are overlapping partly however there is a clear difference between the locations of the IRT and CTT-based posterior density plots. In draw two, three, and four there is a difference in the level-1 (within person) variance posterior density, but no large difference for the level-2 (between person) variance posterior density. The posterior density plots for the random intercept estimates in all draws are overlapping almost completely, indicating no difference in the estimates of the random intercept between the IRT and CTT-based modeling techniques. The reason that the estimates for the intercept are the same for both models is the rescaling procedure that was used as described in equation Eq. 7. The mean and the variance for plausible values are rescaled to the same scale in order to

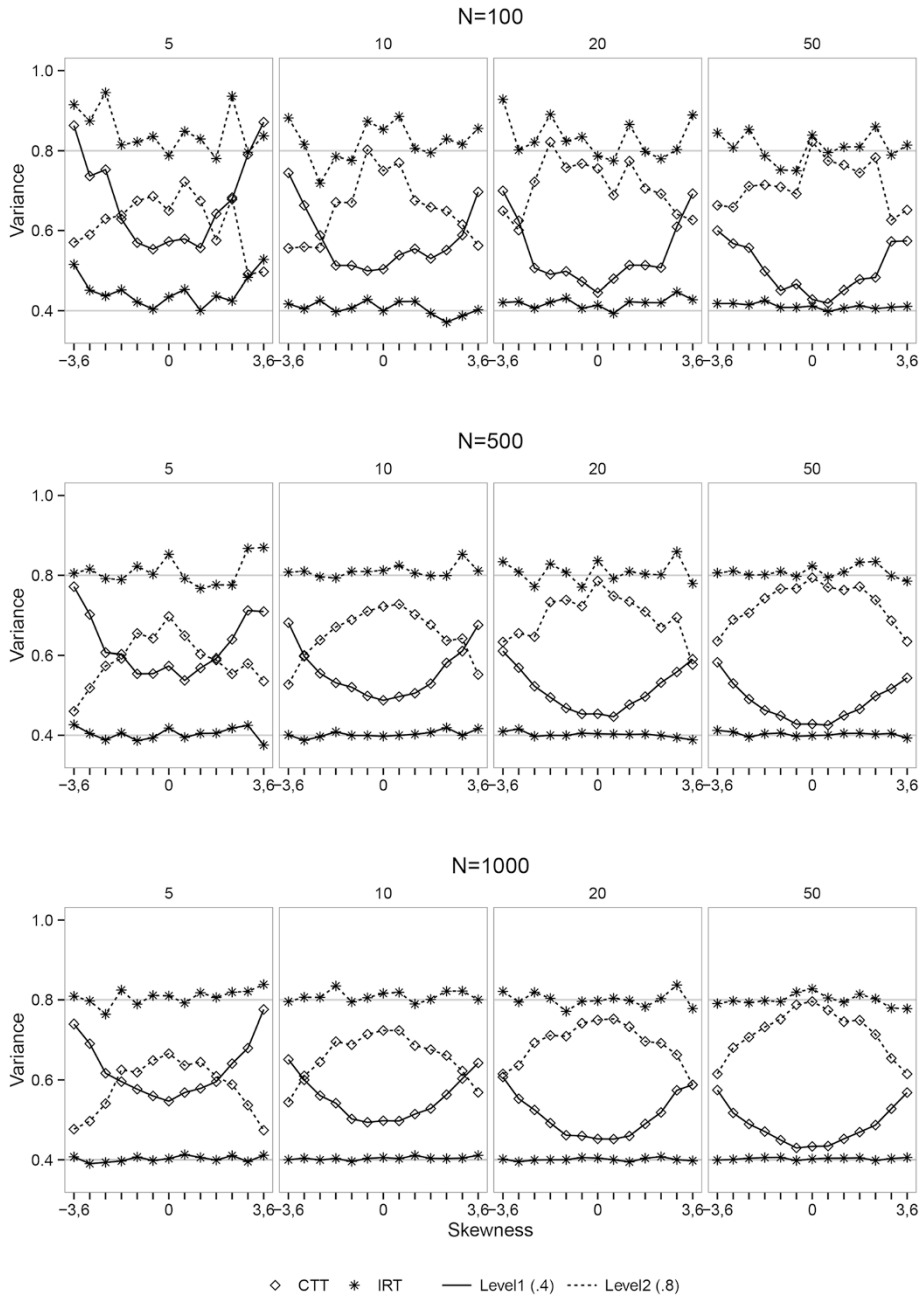

Fig. 4 Pooled variance estimates. Selection of the pooled variance estimates for different distributions, sample sizes ($N = 100, 500, 1000$), and number of items ($K = 5, 10, 20$). The upper left plot for example represents 13 different skewness conditions (from negative to positive) with 100 participants measured on six time points with a five-item questionnaire. The points represent the IRT and CTT-based estimates for the level-1 (repeated measurement) and level-2 (between person) variance. The horizontal lines represent the true values for the variance parameters

guarantee the comparability of the estimates. The results from the data example are in concordance with the results from the simulation study, indicating that the IRT based estimates are closer to the true parameters.

Discussion

Despite the known benefits of IRT modeling when analyzing latent variables, CTT models are still used very often in the field of epidemiological research. The objective of this study was to point out the differences between the use of IRT-based plausible values and the use of sum-scores in the measurement part of a longitudinal analysis. In this study, it is shown that the common way of doing longitudinal analysis with sum-scores leads to systematically biased results and more advanced statistical methods are required to make profound inferences in longitudinal

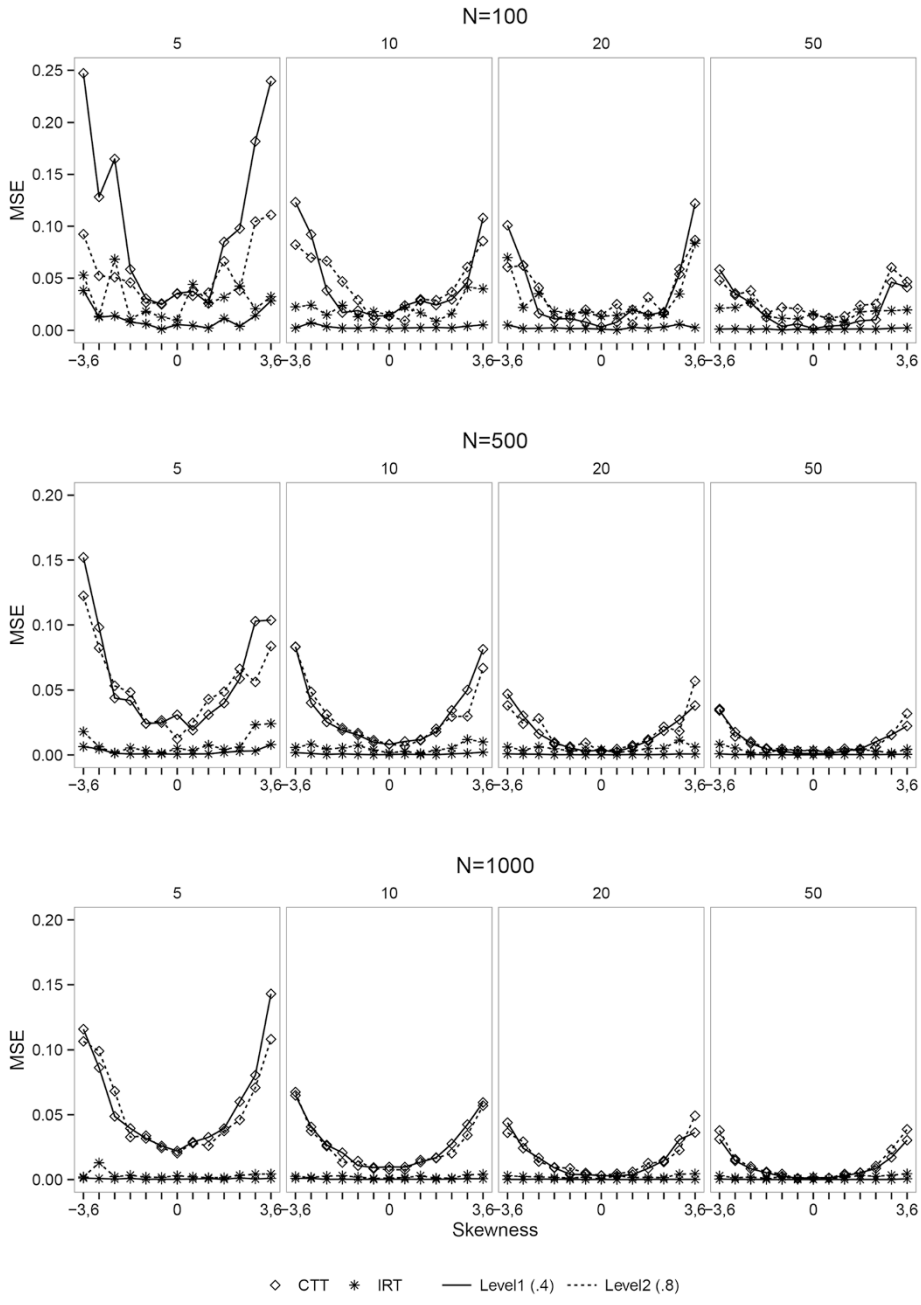

Fig. 5 Mean squared errors. Plots of the MSE's for level-1 (repeated measurement) and level-2 (between person) variance parameter estimates resulting from both the IRT and the CTT-based latent variable models

latent variable research. We showed that IRT-based plausible value techniques performs better compared with CTT analysis for retrieving variance estimates in longitudinal data with latent outcome variables measured with questionnaires. The difference between both methods becomes consistently larger for the more extreme conditions of the simulation, indicating that IRT-based plausible value techniques are quite robust against more extreme data situations. The bias in the CTT based estimates can be reduced by using a larger number of items, a larger sample size, and by using data following a strictly normal distribution. However, in almost all of the data situations in our simulation study, longitudinal IRT performs much better in retrieving the variance estimates. In practice, epidemiological questionnaire data is seldom normally distributed [52–54], and using IRT-based estimates can improve the quality of the estimates profoundly. The systematic underestimation of the between person variance

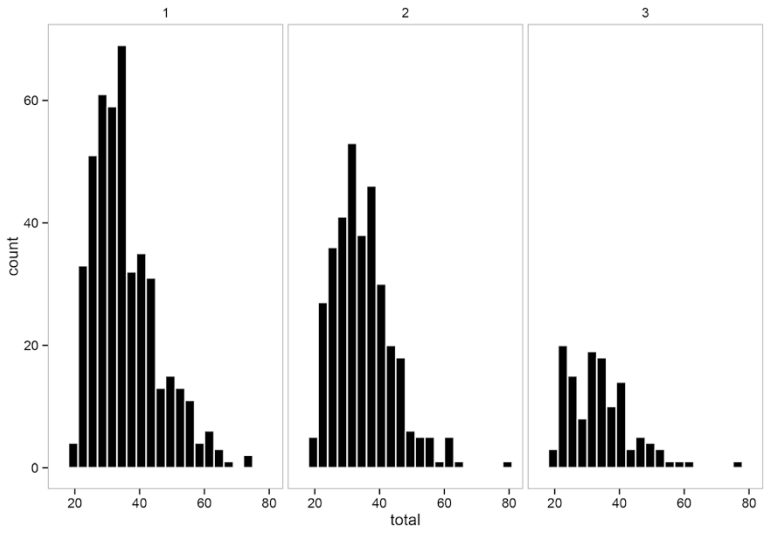

Fig. 6 Sum-score distributions. Histograms with the sum-score distribution for the latent variable 'state anxiety' on the three measurement occasions in the AGHLS cohort. The skewness of the sum-score distributions was 1.02; 0.99; and 1.18 on the first, second, and third measurement occasion from left to right, and 443, 338, and 126 participants were included respectively

and overestimation of the within person variance by the CTT-based model leads to overestimation of the ICC. This might have impact on the regression coefficients and cause bias. It would be interesting to investigate the sequence and direction of this bias and the impact on the conclusions of past and future research. Besides that, nowadays there is also much interest in using multilevel modeling to explain differences between individuals and groups, which makes it even more important to use unbiased estimates for the variance parameters. Based on the outcomes of our research, it is advisable to use IRT-based plausible value techniques when the outcome variable is a repeatedly measured latent variable, especially when the sum-score distribution deviates from strictly normal. Plausible values are not an estimator for the construct, they can never be used to make inferences about individuals. Like most statistical inferences, the objective is to make statements about or comparisons between groups of people.

Table 2 Results data example. Parameter estimates (posterior means) of the multilevel model for the example 'Trait Anxiety' data with a random intercept using the IRT-based plausible values technique and the CTT-based scores as outcome variables

	IRT[a]		CTT[b]	
	Mean[c]	SD	Mean[c]	SD
Fixed effect				
γ Intercept	0.018	0.043	0.015	0.045
Random Effects				
Between individual (level-2)				
τ^2 Intercept	0.733	0.067	0.685	0.074
Within individual (level-1)				
σ^2 Residual variance	0.294	0.030	0.357	0.042
Intra Class Correlation				
ρ	0.287		0.343	

[a]Item response theory based estimates
[b]Classical test theory based estimates using sum-scores
[c]Mean of the coefficients resulting from fitting the structural model (longitudinal multilevel model) to the five draws of plausible values based on the IRT or CTT measurement models

In the work of Blanchin *et al.* [55], longitudinal data modeling results under classical test theory and Rasch IRT models have been compared. In their work, scale-free statistical results are compared as type-I errors and power, since the dependent (latent) variables are not measured on a comparable scale. The CTT and IRT-based analysis showed comparable results in terms of power. This is in contrast to our findings, where we showed a significant increase in bias under the CTT model. However, their comparison is more complex since a common test approach is used (i.e., *t*-test and F-test), which is based on different assumptions in the different modeling approaches. The accuracy of the approximation of the distribution of the test statistic is likely to vary over techniques and models, which could influence the statistical results. Furthermore, differences in estimation methods and modeling differences also influenced their results.

The current study was confined to latent variables measured using questionnaires with items containing four ordinal answering categories. Although this is a common situation, questionnaires with a dichotomous response format (i.e. two answering categories) are used as well for measuring latent variables. When using questionnaires with dichotomous response format there will be less variance in

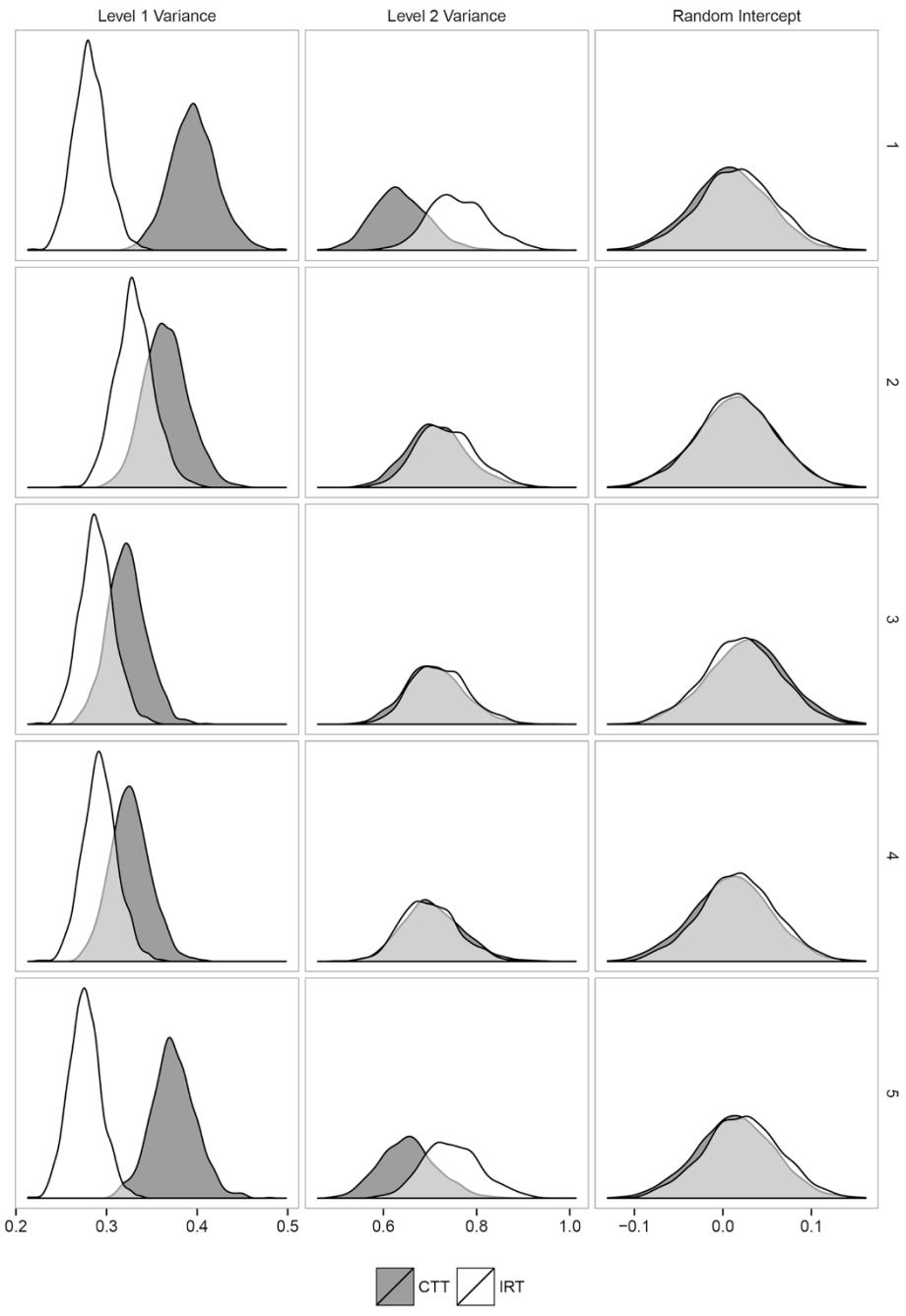

Fig. 7 Posterior density plots. Posterior density plots for the level-1 (repeated measurement) variance, the level-2 (between person) variance, and the random intercept under the IRT and the CTT-based models for all five draws of plausible values

sum-scores compared to ordinal response formats. There are less possible response patterns leading to less variance in scores when aggregated into a sum-score. As a result, the difference in estimates between IRT and CTT will most likely become even larger in all situations.

The simulation study as presented, only took into account complete datasets. Further research is needed to explore the influence of missing data on the difference between both methods. Furthermore, the focus of the current article is the use of latent variables as outcomes in the structural model. Another interesting study would

be to focus on the influence of using CTT based scores for time (in)variant covariates which is a common situation in epidemiological research [56].

Conclusions

From this study it can be concluded that the use of IRT-based latent variable scores, in contrast to sum scores, leads to unbiased parameter estimates in longitudinal data analysis given multi-item questionnaire data. The degree of bias increases when the latent variable distribution is more skewed. It is important to realize that longitudinal data analysis results are biased when using sum scores.

Additional files

Additional file 1: Variance estimates. Contains a table with the IRT and CTT-based pooled variance estimates of the structural model resulting from all different condition of the simulation study. (XLSX 28 kb)

Additional file 2: Mean squared errors. Contains a table with the IRT and CTT-based mean squared errors for the variance estimates of the structural model resulting from all different condition of the simulation study. (XLSX 20 kb)

Abbreviations

IRT: Item Response Theory; AGHLS: Amsterdam Growth and Health Longitudinal Study; CTT: Classical Test Theory; MSE: Mean Squared Error; STAI-DY: State Trait Anxiety Index Dutch Y-version.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JWRT conceived the idea for the comparison aiming to use the best possible methodology for longitudinal data analysis. Also, as a member of the AGHLS board JWRT provided the data that was obtained from the AGHLS cohort. J-PF developed the key idea to use the plausible value imputation technique to make a direct comparison of different measurement methods possible. He developed and adjusted the software that was used for the simulation study and for the application on the empirical data. RG carried out the simulation study and the data analyses using J-PF's software and following his instructions. RG wrote the first version of the manuscript combining the ideas of both J-PF and JWRT on the content and structure of the paper. All authors have read and approved the final manuscript and have provided critical revisions for important intellectual content.

Acknowledgements

We want to thank the Amsterdam Growth and Health Longitudinal Study board and participants for providing us the data and allowing us to illustrate our research with a real life empirical data example. This research was funded by the EMGO+ institute for health and care research.

Author details

[1]Department of Epidemiology & Biostatistics, VU university medical center, Amsterdam, Netherlands. [2]EMGO+ institute for health and care research, Amsterdam, Netherlands. [3]Department of Research Methodology, Measurement, and Data Analysis, Faculty of Behavioral, Management & Social Sciences, University of Twente, Enschede, Netherlands.

Received: 22 January 2015 Accepted: 13 July 2015
Published online: 30 July 2015

References

1. Lin F-J, Pickard A, Krishnan J, Joo M, Au D, Carson S, et al. Measuring health-related quality of life in chronic obstructive pulmonary disease: properties of the EQ-5D-5 L and PROMIS-43 short form. BMC Med Res Methodol. 2014;14:78.
2. Marrero D, Pan Q, Barrett-Connor E, de Groot M, Zhang P, Percy C, et al. Impact of diagnosis of diabetes on health-related quality of life among high risk individuals: the Diabetes Prevention Program outcomes study. Qual Life Res. 2014;23:75–88.
3. Pronk M, Deeg D, Smits C, Twisk J, van Tilburg T, Festen J, et al. Hearing loss in older persons: does the rate of decline affect psychosocial health? J Aging Health. 2014;26:703–23.
4. Bryk A, Raudenbush S. Application of hierarchical linear models to assessing change. Psychol Bull 1987;101:147–58.
5. Kreft I, de Leeuw J, van der Leeden R. Review: review of five multilevel analysis programs: BMDP-5 V, GENMOD, HLM, ML3, VARCL. Am Stat. 1994;48:324–35.
6. Goldstein H. Multilevel Models in Education and Social Research. Oxford: University Press; 1987.
7. Twisk J. Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide. Cambridge: University Press; 2013.
8. Twisk J. Applied Multilevel Analysis: A Practical Guide for Medical Researchers. Cambridge: University Press; 2006.
9. Tuerlinckx F, Rijmen F, Verbeke G, De Boeck P. Statistical inference in generalized linear mixed models: a review. Br J Math Stat Psychol. 2006;59:225–55.
10. Kim J-H, Lee W-Y, Hong Y-P, Ryu W-S, Lee K, Lee W-S, et al. Psychometric properties of a short self-reported measure of medication adherence among patients with hypertension treated in a busy clinical setting in Korea. J Epidemiol. 2014;24:132–40.
11. Golubic R, May A, Benjaminsen Borch K, Overvad K, Charles M-A, Diaz M, et al. Validity of electronically administered recent physical activity questionnaire (RPAQ) in ten European countries. PLoS One. 2014;9, e92829.
12. Leach L, Olesen S, Butterworth P, Poyser C. New fatherhood and psychological distress: a longitudinal study of Australian men. Am J Epidemiol. 2014;180:582–9.
13. Najman J, Khatun M, Mamun A, Clavarino A, Williams G, Scott J, et al. Does depression experienced by mothers leads to a decline in marital quality: a 21-year longitudinal study. Soc Psychiatry Psychiatr Epidemiol. 2014;49:121–32.
14. Astell-Burt T, Mitchell R, Hartig T. The association between green space and mental health varies across the lifecourse. A longitudinal study. J Epidemiol Community Health. 2014;68:578–83.
15. Jarvik J, Comstock B, Heagerty P, Turner J, Sullivan S, Shi X, et al. Back pain in seniors: the Back pain Outcomes using Longitudinal Data (BOLD) cohort baseline data. BMC Musculoskelet Disord. 2014;15:134.
16. Fox J-P. Multilevel IRT using dichotomous and polytomous response data. Br J Math Stat Psychol. 2005;58(1):145–72.
17. Fox J-P. Bayesian modeling of measurement error in predictor variables using item response theory. Psychometrika. 2003;68:169–91.
18. Von Davier M, Gonzalez E, Mislevy R. What are plausible values and why are they useful? IERI Monogr Ser. 2009;9–36.
19. Glas C, Geerlings H, van de Laar M, Taal E. Analysis of longitudinal randomized clinical trials using item response models. Contemp Clin Trials. 2009;30:158–70.
20. Mislevy R. Randomization-based inference about latent variables from complex samples. Psychometrika. 1991;56:177–96.
21. Martin M, Mullis I. TIMSS and PIRLS Achievement Scaling Methodology. In: Methods and procedures in TIMMS and PIRLS. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College; 2011. p. 1–11.
22. Rubin D, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. J Am Stat Assoc. 1986;81:366–74.
23. Wijnstok N, Hoekstra T, van Mechelen W, Kemper H, Twisk J. Cohort profile: the Amsterdam growth and health longitudinal study. Int J Epidemiol. 2012;42:1–8.
24. Lord F, Novick M, Birnbaum A. Statistical Theories of Mental Test Scores. Addison-Wesley Publishing Company, Inc.; 1968
25. Van Nispen R, Knol D, Neve H, van Rens G. A multilevel item response theory model was investigated for longitudinal vision-related quality-of-life data. J Clin Epidemiol. 2010;63:321–30.

26. Van Nispen R, Knol D, Langelaan M, de Boer M, Terwee C, van Rens G. Applying multilevel item response theory to vision-related quality of life in Dutch visually impaired elderly. Optom Vis Sci. 2007;84:710–20.
27. Fox J-P, Glas C. Bayesian modification indices for IRT models. Stat Neerl. 2005;59:95–106.
28. Verhagen J, Fox J-P. Longitudinal measurement in health-related surveys. A Bayesian joint growth model for multivariate ordinal responses. Stat Med. 2013;32:2988–3005.
29. Reise S. Using multilevel logistic regression to evaluate person-Fit in IRT models probability trait level. Multivariate Behav Res. 2000;35:543–68.
30. Hays R, Morales L, Reise S. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38(9 Suppl):1128.
31. Fox J-P. Multilevel IRT Modeling in Practice with the package mlirt. J Stat Softw. 2007;20(5):1–16.
32. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychom Monogr Suppl. 1969;34:100.
33. Embretson S, Reise S. Item Response Theory for Psychologists. L.Erbaum Associates; 2000
34. Pastor D. Longitudinal rasch modeling in the context of psychotherapy outcomes assessment. Appl Psychol Meas. 2006;30:100–20.
35. Bayes T. An essay towards solving a problem in the doctrine of chances. Philos Trans R Soc. 1763;53:370–418.
36. Asparouhov T, Muthén B. Plausible values for latent variables using Mplus. 2010.
37. Rubin D. The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation. J Am Stat Assoc. 1987;82:543–6.
38. Little R, Rubin D. Statistical Analysis with Missing Data. Whiley & Sons; 2002
39. Kolen M, Brennan R. Test Equating, Scaling, and Linking. Methods and Practices. 2nd edition. Springer; 2010.
40. Team R. R: A language and environment for statistical computing. 2012.
41. Sturtz S, Gelman A, Ligges U. R2WinBUGS : a package for running WinBUGS. J Stat Softw. 2005;12:1–16.
42. Lunn D, Thomas A, Best N, Spiegelhalter D. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput. 2000;10:325–37.
43. Kemper H, Hof M van't. Design of a multiple longitudinal study of growth and health in teenagers. Eur J Pediatr. 1978;155:147–55.
44. Hoekstra T, Barbosa-leiker C, Koppes L, Twisk J. Developmental trajectories of body mass index throughout the life course : an application of Latent Class Growth (Mixture) Modelling. Longitunal Life Course Stud. 2011;2(3):319-30.
45. Douw L, Nieboer D, van Dijk B, Stam C, Twisk J. A healthy brain in a healthy body: brain network correlates of physical and mental fitness. PLoS One. 2014;9, e88202.
46. Wijnstok N, Hoekstra T, Eringa E, Smulders Y, Twisk J, Serne E. The relationship of body fatness and body fat distribution with microvascular recruitment: The Amsterdam Growth and Health Longitudinal Study. Microcirculation. 2012;19:273–9.
47. Wijnstok N, Serné E, Hoekstra T, Schouten F, Smulders Y, Twisk J. The relationship between 30-year developmental patterns of body fat and body fat distribution and its vascular properties: the Amsterdam Growth and Health Longitudinal Study. Nutr Diabetes. 2013;3, e90.
48. Twisk J, Kemper H, van Mechelen W, Post G. Tracking of risk factors for coronary heart disease over a 14-year period: a comparison between lifestyle and biologic risk factors with data from the Amsterdam growth and health study. Am J Epidemiol. 1997;145:888–98.
49. Twisk J, Staal B, Brinkman M, Kemper H, van Mechelen W. Tracking of lung function parameters and the longitudinal relationship with lifestyle. Eur Respir J. 1998;12:627–34.
50. Hoekstra T, Barbosa-Leiker C, Twisk J. Vital exhaustion and markers of low-grade inflammation in healthy adults: the Amsterdam Growth and Health Longitudinal Study. Stress Heal. 2013;29:392–400.
51. Van der Ploeg H. De Zelf-Beoordelings Vragenlijst angst (STAY-DY). Tijdschr Psychiatr. 1982;24:189–99.
52. King M, Bell M, Costa D, Butow P, Oh B. The Quality of Life Questionnaire Core 30 (QLQ-C30) and Functional Assessment of Cancer-General (FACT-G) differ in responsiveness, relative efficiency, and therefore required sample size. J Clin Epidemiol. 2014;67:100–7.
53. Hertzog C, van Alstine J. Measurement properties of the Center for Epidemiological Studies Depression Scale (CES-D) in older populations. Psychol assessmen a J Consult Clin Psychol. 1990;2:64–72.
54. Dawson J, Linsell L, Zondervan K, Rose P, Randall T, Carr A, et al. Epidemiology of hip and knee pain and its impact on overall health status in older adults. Rheumatology. 2004;43:497–504.
55. Blanchin M, Hardouin J, Le Neel T, Kubis G, Blanchard C, Mirallié E, et al. Comparison of CTT and Rasch-based approaches for the analysis of longitudinal Patient Reported Outcomes. Stat Med. 2011;30:825–38.
56. Zhang S, Paul J, Nantha-Aree M, Buckley N, Shahzad U, DeBeer J, et al. Empirical comparison of four baseline covariate adjustment methods in analysis of continuous outcomes in randomized controlled trials. Clin Epidemiol. 2014;6:227–35.